INDIXE: EL AGREGADOR NACIONAL DE LA RED MEXICANA DE REPOSITORIOS INSTITUCIONALES-REMERI PARA LA INTEROPERABILIDAD CON REDES FEDERADAS DE REPOSITORIOS INSTITUCIONALES

Rosalina Vázquez Tapia[1], *Antonio Felipe Razo Rodríguez* [2]

[1]Maestra en Tecnología Educativa; Directora de la Biblioteca Virtual Universitaria, Universidad Autónoma de San Luis Potosí, Niño Artillero S/N Zona Universitaria, San Luis Potosí, S.L.P., México. alinavn@uaslp.mx

[2]Maestro en Diseño de la Información; Profesor del Departamento de Arte, Diseño y Arquitectura. Universidad Iberoamericana Puebla. Blvd. del Niño Poblano No. 2901 Unidad Territorial Atlixcáyotl, Puebla, Pue, México. antrazo@gmail.com

RESUMEN

La Red Mexicana de Repositorios Institucionales (REMERI), tiene como objetivo crear una red interconectada de repositorios digitales de instituciones de educación superior (IES) en México, para integrar, difundir, preservar y dar visibilidad a su producción científica, académica y documental; así como también, incorporarse a redes o directorios de repositorios internacionales para fomentar la colaboración y apoyar el acceso y la divulgación de contenidos de acceso abierto. El proyecto REMERI (www.remeri.org.mx) inicia en el 2011 a partir de una iniciativa presentada por la Biblioteca Virtual de la Universidad Autónoma de San Luis Potosí, para desarrollar una Red de Repositorios Institucionales de las instituciones miembros de la Red Abierta de Bibliotecas Digitales (RABID), perteneciente a la Corporación Universitaria para el Desarrollo de Internet 2 (CUDI). Se integra un grupo de trabajo de seis instituciones mexicanas de educación superior miembros de RABID, mismas que se constituirían posteriormente como fundadoras de REMERI. Ante la necesidad de contar con una plataforma tecnológica que permitiera la creación de un nodo mexicano interoperable con la Red Federada de Repositorios Institucionales de Publicaciones Científicas LA-Referencia, los representantes por parte de México en dicho proyecto, adoptan la iniciativa de REMERI para construir un primer prototipo con base en estándares internacionales predefinidos. Para la interconexión de los repositorios institucionales y su interoperabilidad con LA-Referencia, fue necesario considerar el uso de estándares de interoperabilidad basados en DRIVER 2.0 y la implementación de servidores de metadatos con el protocolo OAI-PMH. En un primer análisis de los repositorios disponibles en México para la construcción del prototipo, se encontraron servidores no estandarizados, metadatos inconsistentes, faltantes o no estructurados y una implementación parcial o con modificaciones del protocolo OAI-PMH; además, de una diversidad de recursos, formatos y áreas de conocimiento. Adicionalmente, se establecieron los siguientes requerimientos para el desarrollo de la plataforma tecnológica de REMERI: Integración en un mismo sistema de diversas colecciones de Repositorios Institucionales; recuperación de información por medio de una interfaz de búsqueda y consulta web; indexación de la colección para mostrar resultados por relevancia; interoperabilidad de la colección

utilizando el protocolo OAI-PMH; uso y normalización de los metadatos a estándares y especificaciones Dublin Core, LA Referencia y DRIVER. El resultado fue el desarrollo de un *cosechador agregador* y sistema de consulta denominado INDIXE, el cual tiene las siguientes características: desarrollo propio basado en el uso de software *open source*, configurable y flexible; desarrollo funcional con tolerancia a errores y fácil de administrar; diseño de una interfaz accesible vía web multiplataforma y navegable para la incorporación, mantenimiento, consulta e interoperabilidad de repositorios institucionales; acceso a las colecciones de REMERI por medio de mecanismos avanzados de recopilación, búsqueda y visualización de grandes colecciones; implementación de estándares web para garantizar la usabilidad en plataformas fijas y móviles, implementación de estándares de metadatos y normalización de información para facilitar su interoperabilidad. En esta propuesta se describen los componentes y arquitectura de la plataforma INDIXE, sus servicios y posibilidades de crecimiento; así como el trabajo efectuado para lograr la interoperabilidad con LA-Referencia.

Palabras clave: Interoperabilidad, Acceso Abierto, Repositorios Institucionales.

INTRODUCCIÓN

El creciente desarrollo y la rápida diseminación de la información y del conocimiento, junto con una serie de fenómenos globales de carácter económico, social y tecnológico, propiciaron hace más de una década, el surgimiento de un nuevo paradigma de la comunicación científica denominado *Acceso Abierto* (AA), el cual promueve a partir de una serie de declaraciones y definiciones, que el conocimiento científico debe ser de acceso libre y gratuito a todo mundo, sin barreras económicas, políticas, legales o de cualquier otra índole.

El movimiento de Acceso Abierto

El concepto de AA para datos científicos, se utilizó por primera vez para compartir la colección de datos referentes al año Internacional de Geofísica (International Council of Science, 1957-1958). En el 2002, la Iniciativa de Acceso Abierto de Budapest [Budapest Open Access Initiative, (BOAI)] lanzó una campaña mundial para el AA, también llamado *Open Access*, (OA) para las publicaciones científicas. La iniciativa BOAI fue la primera en utilizar el concepto de AA y la primera en proponer las dos estrategias del AA: las revistas (llamada vía dorada) y los repositorios (llamada vía verde). En Junio de 2003, en la Declaración de Bethesda (Bethesda Statement on Open Access Publishing), surgieron la primera definición de las publicaciones de AA y el concepto de depósito o autoarchivo en repositorios. En octubre del mismo año, en la Declaración de Berlín sobre Acceso Abierto al Conocimiento en Ciencias y Humanidades (Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities) se definieron las contribuciones del AA y las dos condiciones que éstas deben cumplir.

Entre las principales definiciones del AA destacan las tres siguientes:

1. "Por *acceso abierto* a la literatura científica erudita, entendemos su disponibilidad gratuita en Internet, para que cualquier usuario la pueda leer, descargar, copiar, distribuir o imprimir, con la posibilidad de buscar o enlazar al texto completo del artículo, recorrerlo para una indexación exhaustiva, usarlo como datos para software, o utilizarlo para cualquier otro propósito legal, sin otras barreras financieras, legales o técnicas distintas de la fundamental de acceder a la propia Internet. El único límite a la reproducción y distribución de los artículos publicados, y la única función del copyright en este marco, no puede ser otra que garantizar a los autores el control sobre la integridad de su trabajo y el derecho a ser acreditados y citados*"* (Budapest Open Access Initiative, 2003, p. 1).
2. En la declaración Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003) se establece la siguiente definición de AA: Definimos el Acceso Abierto como una amplia fuente de conocimiento humano y patrimonio cultural aprobada por la comunidad científica. Las contribuciones del acceso abierto incluyen los resultados de la investigación científica original datos primarios y metadatos, materiales fuentes, representaciones digitales de materiales gráficos y pictóricos, y materiales eruditos en multimedia.
3. La literatura de AA es digital, en línea, libre de cargos y libre de la mayoría de las restricciones de *copytight* y licencias (Suber, 2012).

De acuerdo a Suber (2012), hay muchas formas de hacer posible el AA: páginas web personales, *blogs*, *wikis*, foros de discusión, libros electrónicos, bases de datos, *webcasts*, recursos multimedia, sindicación de contenidos RSS, páginas de noticias, entre otros. Sin embargo, los dos vehículos o rutas de AA más implementados son los repositorios y las revistas.

El movimiento Open Access (OA), utiliza el término Gold OA (AA Dorado) para la entrega a través de las Revistas; y Green OA (AA Verde) para la entrega a través de Repositorios. El concepto Self-Archiving (Autoarchivo) es la práctica de depositar por uno mismo un trabajo en un repositorio OA. Los tres términos fueron acuñados por Steven Harnard, (Suber, 2012).

Definiciones de Repositorio Institucional

El archivo de contenidos científicos en repositorios digitales constituye la vía verde del AA. En la década de los 90´s se usaban los archivos de eprints (*eprint archives*) y archivos abiertos (*open archives*) para hacer referencia al archivo electrónico de contenidos científicos. Posteriormente con el movimiento de AA, se utilizó el término repositorio como sinónimo de archivo, el cual se consolidó y desplazo a los conceptos anteriores. Un repositorio digital puede definirse como un sistema en red que proporciona servicios Web sobre una colección de objetos digitales, basado en una arquitectura abierta y en el uso de estándares (Abadal, 2012).

De acuerdo a su propósito, los repositorios digitales se clasifican en dos tipos: los institucionales y los temáticos. Los Repositorios Institucionales (RI) son aquellos que almacenan, preservan y proveen acceso a la producción científica y académica de una institución. Los repositorios temáticos albergan colecciones o recursos de una determinada disciplina o área de conocimiento específica; pueden ser creados o dirigidos por instituciones académicas, de investigación y también por organismos gubernamentales. Existen otros tipos de repositorios en función de su alcance, características y tipo de contenidos.

Uno de los mayores precursores y promotores del desarrollo de RI es Clifford Lynch, Director Ejecutivo de la Coalition for Networked Information, quien en el año 2003 proporcionó una de las primeras definiciones: Un Repositorio Institucional es un conjunto de servicios que la universidad ofrece a los miembros de su comunidad para la gestión y para la diseminación de los materiales en forma digital creados por la institución y sus miembros.

En el nivel más básico y fundamental, un RI es un reconocimiento a la vida intelectual y académica de una institución que será representada, documentada y compartida en forma digital; y para ello, una responsabilidad primordial de las universidades es implementar mecanismos para el acceso y preservación de los contenidos (Lynch, 2003). Un Repositorio Institucional "es un archivo digital de la producción intelectual creada por una facultad, un equipo de investigación, y estudiantes de una institución y accesible a los usuarios finales dentro y fuera de la institución, con pocas barreras de acceso o ninguna" (Crow, 2002, p. 17).

Además de los Repositorios Institucionales y temáticos, se identifican otros siete tipos de repositorios: a) Huérfanos, creados para el depósito de trabajos de autores o investigadores que no tienen acceso a otro repositorio institucional o temático; b) de Datos, que almacenan y preservan los datos científicos asociados a un proyecto de investigación; c) Centralizados o especializados, creados por las sociedades científicas, las asociaciones, las entidades gubernamentales o los organismos de financiamiento; d) Multi-institucionales, que integran la producción intelectual de un conjunto de instituciones pertenecientes a una red o consorcio a nivel regional o nacional; e) Integradores o agregadores, son recolectores, agregadores o portales que recolectan los contenidos de varios repositorios institucionales o temáticos; f) de Revistas arbitradas, que almacenan y preservan los titulos de revistas científicas publicadas o financiadas por una determinada institución; g) por tipo de contenidos, son repositorios que almacenan solamente determinado tipo de documentos.

Respecto al tipo de contenidos que pueden albergar los RI, Alonso, Subirats y Martínez (2008) los clasifican en tres tipos: productos científicos, productos institucionales o administrativos y objetos de aprendizaje. Adicionalmente, algunos repositorios institucionales, sobre todo los de España, incluyen acervos antiguos o documentos patrimoniales en sus colecciones.

Es importante aclarar que un RI no es un canal de publicación, sino una vía de comunicación científica, de difusión y visibilidad del conocimiento y debe comprenderse como complementario al proceso de publicación científica formalizado con revisión por pares (Bustos y Fernández, 2008).

Soluciones de software y servicios de metadatos asociados a los repositorios

Existen diferentes aplicaciones de software de código abierto (open source) para crear repositorios institucionales o temáticos, entre los más utilizados se encuentran DSpace, EPrints, Fedora e Islandora. También existen programas comerciales, como Digital Commons, CONTENTdm, DigiTool y EQUELLA Repository.

Para la recuperación del contenido de los repositorios digitales es necesario implementar servicios de metadatos a través de protocolos de intercambio de información. El estándar Dublin Core es uno de los esquemas de metadatos más utilizados que define 15 elementos básicos para describir un documento. También existen estándares para describir recursos más complejos o una agrupación de recursos, como los estándares METS o MODS. Para los objetos de aprendizaje se utilizan el SCORM y el LOM. Para las tesis digitales se utiliza el estándar ETD-MS.

Por lo general, los servicios de metadatos implementan el protocolo estándar de recolección de archivos abiertos, denominado OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). El protocolo es un servicio Web que permite recuperar la información del repositorio, sus colecciones, los identificadores de los recursos y los metadatos de los recursos.

El proceso de recuperación de metadatos se conoce como recolección o cosecha (*harvesting*). Estos registros se almacenan en una base de datos o en un indexador con una estructura determinada, los cuales pueden ser consultados a través de una interfaz Web.

Algunos de los programas libres disponibles para la creación de integradores o recolectores son DSpace, VuFind, Open Harvester System OHS de PKP y el programa de D-Net del proyecto DRIVER, que es específicamente para infraestructuras complejas.

Existen servicios que validan y certifican los servidores de metadatos de acuerdo al estándar OAI-PMH. Estos servicios se denominan validadores y permiten evaluar y garantizar la interoperabilidad de los repositorios institucionales. Para ello, es necesario utilizar vocabularios controlados o especificaciones de interoperabilidad como DRIVER 2.0 y OpenAIRE, en Europa.

LA RED MEXICANA DE REPOSITORIOS INSTITUCIONALES-REMERI

La Red Mexicana de Repositorios Institucionales es una red federada de repositorios institucionales y temáticos de las Instituciones Mexicanas de Educación Superior y de Investigación, que recolecta e integra su producción científica, académica y documental, para su difusión, visibilidad y acceso abierto. Se puede definir también como un Repositorio agregador de alcance nacional que contiene las referencias (metadatos) de repositorios digitales y revistas de acceso abierto de México.

Antecedentes y desarrollo

El proyecto REMERI surge en principio de la necesidad de contar con una plataforma para integrar los repositorios digitales de las Instituciones Mexicanas de Educación Superior que permitiera su difusión, localización y visualización de manera interoperable a través de interfaces comunes.

Con este propósito, en el 2011 la Biblioteca Virtual de la Universidad Autónoma de San Luis Potosí (UASLP), presenta una iniciativa para desarrollar una Red de Repositorios Institucionales de las instituciones miembros de la Red Abierta de Bibliotecas Digitales (RABID), perteneciente a la comunidad de bibliotecas digitales de la Corporación Universitaria para el Desarrollo de Internet 2 (CUDI).

A partir de esta propuesta, se integra un grupo de trabajo de seis instituciones miembros de RABID, mismas que se constituirían posteriormente como fundadoras de REMERI: Universidad de las Américas Puebla (UDLAP), Instituto Tecnológico de Estudios Superiores de Monterrey (ITESM), Universidad de Guadalajara (UDG), Universidad Autónoma de San Luis Potosí (UASLP), Universidad Autónoma del Estado de Hidalgo (UAEH) y Universidad Autónoma del Estado de México (UAEM).

Por otro lado, en octubre de 2011 se lleva a cabo en la ciudad de México, D.F., la 4º Reunión de trabajo de miembros del proyecto impulsado por la RedCLARA y financiado por el Banco Interamericano de Desarrollo denominado "*Red Federada de Repositorios Institucionales de Publicaciones Científicas LA-Referencia"*. En este proyecto participan representantes de 9 países de Latinoamérica: Brasil, México, Argentina, Chile, Colombia, Perú, Ecuador, Venezuela y El Salvador.

En el marco de esta reunión y ante la necesidad de contar con una plataforma tecnológica que permitiera la creación de un nodo mexicano interoperable con LA-Referencia, los representantes por parte de México en dicho proyecto, el Consejo Nacional de Ciencia y Tecnología (CONACyT) y CUDI, adoptan la iniciativa de REMERI para construir un primer prototipo con base en estándares internacionales predefinidos.

De esta manera, en noviembre del 2011 bajo el liderazgo de la UASLP, el apoyo de CUDI y el financiamiento de CONACyT, el grupo de trabajo de las instituciones fundadoras de REMERI, presentan el proyecto con el propósito de crear una Red federada de repositorios digitales de acceso abierto y la implementación de un primer prototipo nacional para LA-Referencia, designando a la UASLP como el responsable técnico del nodo mexicano en LA- Referencia.

Durante el 2012 se lleva a cabo la primera fase que comprendió el desarrollo de seis componentes estratégicos. Como parte de los resultados, se diseñaron documentos normativos para la operación de la Red, los requisitos técnicos de adhesión y modelos de sostenibilidad financiera; se llevo a cabo un diagnóstico sobre el desarrollo de repositorios institucionales sobre una muestra representativa de 55 instituciones; se organizaron presentaciones y talleres de difusión y capacitación tanto virtuales como presenciales; y finalmente, se implementó la plataforma tecnológica de REMERI y una primer piloto interoperable con LA-Referencia, con 6 Repositorios Institucionales y 3,696 registros indexados.

En abril de 2013, con el apoyo de CUDI y bajo la coordinación general de la UASLP se conforma un grupo técnico de soporte para implementar la segunda fase de REMERI con cuatro objetivos fundamentales: Incorporar a más instituciones y repositorios a la Red y al nodo de LA-Referencia; capacitar y brindar apoyo a las instituciones que no contaban con repositorios estandarizados; consolidar el desarrollo de la plataforma tecnológica; y, formalizar la estructura y administración de la Red para su sostenibilidad operativa y financiera a largo plazo.

Durante este segundo año, a través de estrategias de asesoría y análisis y normalización de metadatos, se logró la integración y recolección de 53 repositorios institucionales y temáticos pertenecientes a 27 Instituciones Mexicanas de Educación Superior con un total de 179,546 documentos. Además, se implementaron mejoras al sitio web y nuevos servicios y herramientas.

Posteriormente, en noviembre de 2013 CUDI convoca a una reunión de miembros fundadores para establecer el modelo de gobernanza de la Red, mismo que es aprobado por su Consejo Administrativo en Febrero de 2014, constituyéndose a REMERI como uno de sus proyectos estratégicos y nueva comunidad de CUDI.

Actualmente (agosto de 2014), REMERI cuenta con 69 repositorios de 37 instituciones para un total 208,695 registros de metadatos de tesis, artículos y libros. De este conjunto, casi 100,000 registros de 30 instituciones están indexadas por LA-Referencia.

Servicios de REMERI

Los servicios que ofrece REMERI son los siguientes:

- Un portal web con noticias, redes sociales, eventos, documentos, informes de incorporación y colecciones, enlaces a servicios, formularios de registro y consultas, directorio de participantes, enlaces a repositorios, servicio de consulta, información de capacitación, material informativo y multimedia. [1]
- Un servicio de búsqueda para todas las colecciones con resultados ordenados y filtrados por relevancia, fecha, institución, autor y tipo de documento. Paginación y segmentado de resultados. Ficha informativa y detallada de los recursos con iconografía, enlace al documento, al repositorio y a sus metadatos. [2]
- Servicio de validación de servidores de metadatos con consulta y validación sintáctica (en base al *Schema*) de la respuesta a los verbos del protocolo *OAI-PMH* con manejo de errores. [3]
- Formulario de registro para Repositorios Institucionales con información de contacto, descripción del contenido, contactos administrativos y técnicos así como enlaces al repositorio y a su servidor de metadatos. El proceso de registro facilita el proceso de análisis y diagnóstico del repositorio para ser considerado para su incorporación. [4]

[1] http://www.remeri.org.mx
[2] http://www.remeri.org.mx/portal/REMERI.jsp?busca=tesis
[3] http://www.remeri.org.mx/portal/valida.html
[4] http://www.remeri.org.mx/portal/formulario-registro.html

- Una vez cosechado, normalizado e integrado el repositorio, se incorpora al *INDIXE de Repositorios Institucionales* de REMERI, un directorio disponible para consulta desde. [5]
- A su vez el contenido se reporta en el *INDIXE de Producción Científica*, una relación de artículos y tesis de las instituciones de educación superior del país basado en el contenido de sus repositorios. [6]
- Servidores de metadatos institucionales, es un servicio de metadatos de aquellas colecciones y repositorios que no cuentan con los mecanismos para implementar el servicio en sus plataformas. [7]
- Servidor de metadatos de la red, es un servicio de metadatos que permite recuperar todas las colecciones y registros de la red. [8]
- Servidor de metadatos de producción científica estandarizado de acuerdo a los requerimientos de LA-Referencia (basados a su vez en *DRIVER*). [9]

AGREGADOR Y SISTEMA DE CONSULTA INDIXE

INDIXE es un servicio cosechador-agregador (integrador) ideal para repositorios temáticos, regionales o nacionales, desarrollado específicamente para el proyecto REMERI. Es al mismo tiempo un proveedor de datos ya que cuenta con un servicio de metadatos que le permiten a la colección integrarse a otras redes.

Requerimientos técnicos de REMERI

Los requerimientos funcionales definidos para el proyecto REMERI fueron los siguientes: integración en un mismo sistema de diversas colecciones de Repositorios Institucionales; recuperación de información por medio de una interfaz de búsqueda y consulta web; indexación de la colección para mostrar resultados por relevancia; interoperabilidad de la colección utilizando el protocolo OAI-PMH; uso y normalización de los metadatos a estándares y especificaciones Dublin Core, LA-Referencia y DRIVER.

Los requerimientos técnicos definidos para el desarrollo de la plataforma tecnológica consideraron lo siguiente: simplificar la instalación y administración de componentes, facilitar la gestión y administración de los servicios de preferencia con interfaz vía web, permitir procesar y rectificar metadatos recolectados, ofrecer además un servidor de metadatos de la información recolectada de acuerdo a DRIVER.

[5] http://www.remeri.org.mx/repositorios
[6] http://www.remeri.org.mx/produccion.
[7] http://www.remeri.org.mx/indixe/rest//db/remeri/servicios/unam/tesis/oai_server_tesis_unam.xq?verb=Identify
[8] http://www.remeri.org.mx/indixe/rest//db/remeri/oai/oai_server.xq?verb=Identify
[9] http://www.remeri.org.mx/indixe/rest//db/remeri/driver/driver_server.xq?verb=Identify

Para el desarrollo e implementación de este proyecto se consideraron las siguientes alternativas :

- OAIConnect. (www.drupal.org/project/oaiconnect). Permite la recolección de metadatos y los almacena en una estructura estándar en una base de datos relacional. Funciona sobre Drupal, Solr, PHP y MySQL con servidor Apache. Se requerían instalar y configurar tres servicios, no consideraba la transformación de metadatos y no contaba con servidor de metadatos.
- DSpace. (www.dspace.org). Además de gestionar contenidos, permite la recolección de metadatos y los almacena en una base de datos relacional. Funciona con Java sobre Apache Tomcat, Solr, PostgreSQL u Oracle con JSP. Se requerían instalar y configurar tres servicios, no consideraba la transformación de metadatos y en su momento no contaba con el estándar DRIVER para su servidor de metadatos.
- D-Net. (www.d-net.research-infrastructures.eu). Proyecto promovido por la Unión Europea en su iniciativa DRIVER. Permite la agregación de registros y metadatos, transformación y limpieza de metadatos y los almacena en una base de datos no relacional (JSON style). Funciona con Java sobre Jetty, MongoDB, Python. Se requerían instalar y configurar tres servicios, no consideraba la transformación de metadatos.
- VuFind (www.vufind.org) plataforma para la gestión y consulta de registros bibliográficos, incluye recolector de metadatos, almacena la información en Solr y funciona con PHP y MySQL con servidor Apache. Se requerían instalar y configurar tres servicios, no consideraba la transformación de metadatos y en su momento no consideraba el estándar DRIVER para su servidor de metadatos.
- Desarrollo propio. Como alternativa se contaba con la posibilidad de desarrollar la solución basándose en los requerimientos y particularidades del proyecto.

Se consideraron estas opciones y se decidió por el desarrollo propio con la plataforma eXist XML-DB de código abierto. Dentro de las consideraciones para su elección se encontraron: que es una plataforma de desarrollo todo en uno (base de datos, administración web y servidor web); el uso de una misma tecnología (XML) para almacenamiento de la información, procesamiento y consulta; indexación del texto de la colección para consultas de documentos por relevancia; el uso del mismo lenguaje para todo el desarrollo (XQuery) que no requiere compilación ya que es interpretado; interfaz de gestión y administración web, que no requiere el acceso a consola del servidor; instalación una sola vez y un único servicio; plataforma web robusta y probada (Java-Tomcat).

Además, en un primer análisis de los repositorios disponibles en México para la construcción del prototipo, se encontraron servidores con metadatos no estandarizados, inconsistentes, faltantes o sin estructura, con implementaciones parciales o con modificaciones del protocolo OAI-PMH; adicionalmente, una diversidad de recursos, formatos y áreas de conocimiento, por lo que en su momento ningún repositorio analizado cumplía con las especificaciones de DRIVER.

Arquitectura y funcionalidad del Sistema INDIXE

Para la cosecha, indexación y consulta de los repositorios agregados a REMERI, se desarrolló una herramienta propia denominada INDIXE basado en tecnologías XML (XQuery, XPath, XSLT). La base de datos utilizada (eXist) almacena y procesa los metadatos en el formato XML que es el formato nativo en la que se encuentran los metadatos (Dublin Core). La programación de tareas, procesos, servicios y consultas se realizaron en el lenguaje XQuery especializado para información semiestructurada. La base de datos cuenta con indexación de la colección con Lucene con un método de espacios vectoriales y booleano. Se consiguió con esto un desarrollo escalable, de alto desempeño, eficiente, con código compacto y multiplataforma. La arquitectura del sistema INDIXE se muestra en la figura 1.

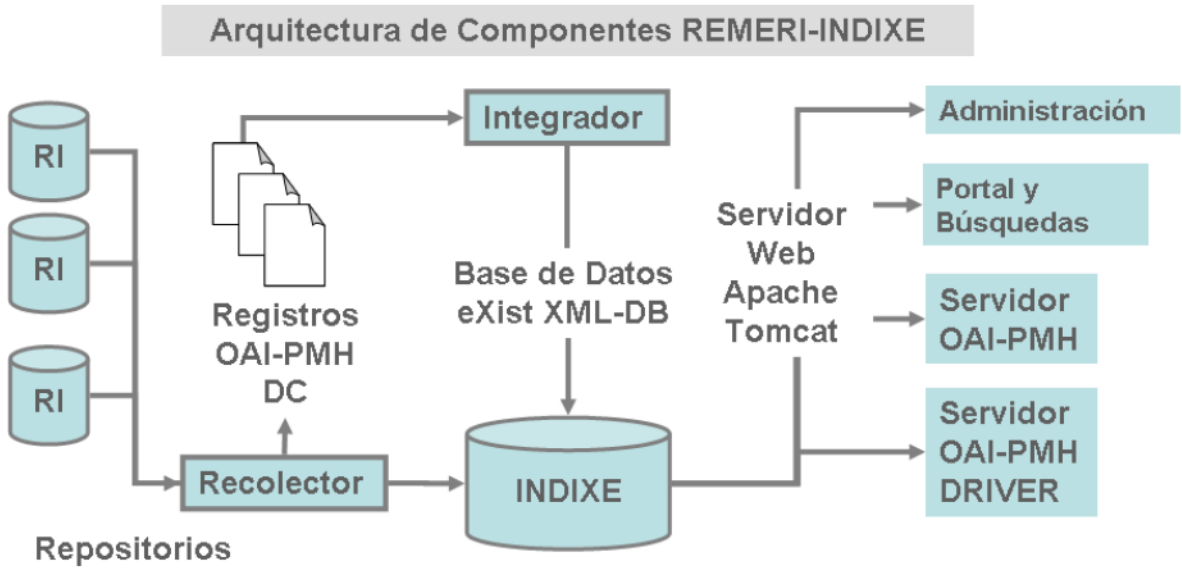

Figura 1. Diagrama de Componentes

Dentro de las aplicaciones o herramientas desarrolladas en el proyecto se encuentran implementadas en el lenguaje XQuery y son consultadas vía web:

- Un validador (sintáctico) para servidores de metadatos
- Un servicio de recolección (cosecha)
- Un servicio integrador (con normalización y estructura de metadatos)
- Un servicio de consulta y recuperación de información (ordenado por relevancia, fecha, institución, tipo o autor)
- Un servidor de metadatos en el estándar OAI-PMH,
- Un servicio integrador para LA-Referencia (DRIVER) con normalización de tipo de documentos y la incorporación del campo de institución (instname) y repositorio (reponame)
- Un servidor de metadatos en el estándar OAI-PMH para LA-Referencia (DRIVER) con especificaciones técnicas requeridas con tipos de documentos y tamaño de la consulta (token).

La infraestructura tecnológica sobre la que opera la plataforma de software consiste en un servidor Dell PowerEdge R720xd con dos procesadores, 64 GB en RAM y 3.5 TB en disco con sistema operativo Linux (OpenSUSE 12.1 amd-64). La plataforma de desarrollo (JSP), el servidor de web y de aplicaciones (Tomcat 6.0) junto la base de datos (eXist 1.4.2) funcionan con Java (OpenJDK 1.6).

INTEROPERABILIDAD DE REMERI CON REDES FEDERADAS DE REPOSITORIOS INSTITUCIONALES DE PUBLICACIONES CIENTÍFICAS

En REMERI pueden participar instituciones de educación superior públicas y privadas, centros de investigación, organizaciones sin fines de lucro y, en general de instituciones de los sectores público y privado en México. Todo miembro de la Red debe configurar su repositorio de forma que permita compartir información de forma homologada, utilizar un estándar de metadatos preferentemente Dublin Core y se hace una fuerte recomendación de cumplimiento con el protocolo OAI-PMH y las directrices DRIVER 2.0. Los metadatos requisito que todo miembro de REMERI debe asegurarse de configurar en su repositorio, son los siguientes:

- dc:title
- dc:identifier
- dc:creator
- dc:type
- dc:date

Los metadatos que en caso de no localizarse pueden ser procesados y asignados de forma automatizada a través del sistema de recolección de metadatos son los siguientes:

- dc:rights
- dc:publisher
- dc:language
- dc:format
- dc:source

Los metadatos que no son requisito, sin embargo se hace una fuerte recomendación que se llenen de forma correcta y se provean, son los siguientes:

- dc:description
- dc:subject
- dc:contributor

El servidor de metadatos del Repositorio Institucional debe aprobar el proceso de validación (sintáctica) y una vez analizado y evaluado su contenido, se establece el proceso de integración. El proceso de recolección requiere la completa implementación de los verbos de OAI-PMH para recuperar los identificadores (ListIdentifiers) y cada registro de manera particular (GetRecord). Las recolecciones se realizan completas cada vez, para que en el caso de hacer ajustes en el proceso de transformación, la colección resulte consistente. El proceso de transformación consiste en mapear, normalizar y estandarizar los metadatos (de acuerdo a los casos presentados anteriormente) y al mismo tiempo, genera la versión de los registros para su consulta e incorporación al servidor de metadatos.

Interoperabilidad con LA- REFERENCIA

Para el caso de las colecciones definidas en LA-Referencia, las tesis de maestría, doctorado, artículos y reportes de investigación (únicos documentos establecidos por al Red para ser cosechados), se cuenta por cada repositorio con un proceso de transformación de acuerdo a los requerimientos de DRIVER para el uso de vocabularios, términos y elementos, a su vez, se incorpora la colección del servidor de metadatos DRIVER para ser incorporados al servidor de metadatos consultado por LA-Referencia.

La experiencia en este proyecto nos ha permitido identificar problemáticas comunes en los diferentes tipos de repositorios, como las siguientes:

dc:identifier

Es común en el caso del software DSpace, exponer el uso del identificador con un handle cuando este no se encuentra activo para la institución y el servidor, también se ha encontrado el uso del IP del servidor o el término “localhost”. En muchos casos los administradores de repositorios no están al tanto de esto y pueden contar con una gran cantidad de registros expuestos de esta manera. Esto se puede corregir en el proceso de integración con INDIXE pero el repositorio original no es consistente para otros proyectos. Es recomendable el acceso al documento directamente, no a páginas de presentación o intermedias, de manera que se agilice la consulta al registro.

dc:type

Es común encontrar el tipo “otros” o registros sin tipos. Es mejor definir en el repositorio a detalle los tipos de materiales para su correcta consulta e incorporación. Para repositorios temáticos es posible asignar los tipos de manera automática. En el caso del proyecto REMERI y LA-Referencia es recomendable agregar o especificar el tipo de tesis, “Tesis de Maestría” o “Tesis de Doctorado” al tipo genérico “Tesis”. En el caso de repositorios con conjuntos (DSpace) esto se puede resolver de manera automatizada en el proceso de integración pero hay que analizar caso por caso. Lo más recomendable es utilizar los vocabularios de DRIVER para tipos de datos.

dc:date
Es común encontrar más de una ocurrencia para las fechas en el caso de plataformas de gestión (se incluye la fecha de registro y la fecha de última actualización). La fecha de publicación generalmente se encuentra en la misma posición y es posible recuperarla con INDIXE de manera automatizada. Existen registros sin fecha o con la fecha en el formato no estándar, también se procesa en la integración nuevamente caso por caso. Lo mas recomendable es mostrar una sola fecha.

dc:publisher
Casi ningún proveedor (por default) proporciona el nombre de la Institución y el repositorio de dónde se está obteniendo la información. Se agrega esta información como metadato a cada registro, no se debe de obviar ya que los metadatos son procesados automáticamente e integrados con otros, el nombre del identificador o URL no siempre es significativo y menos aún cuando se utiliza un IP. Esto también se procesa en la integración en cada caso requerido.

En el caso de LA-Referencia (www.lareferencia.info), REMERI formó parte de las pruebas técnicas desde octubre del 2012. Hasta el momento, México es la red nacional que incorpora la mayor cantidad de registros en idioma Español en el proyecto con un total de 95,194 (en agosto de 2014) provenientes de un total de 55 repositorios institucionales de 31 instituciones, manteniendo un crecimiento continuo y actualizado. Las estadísticas de recolección de México en la referencia por fecha se pueden ver en la figura 2.

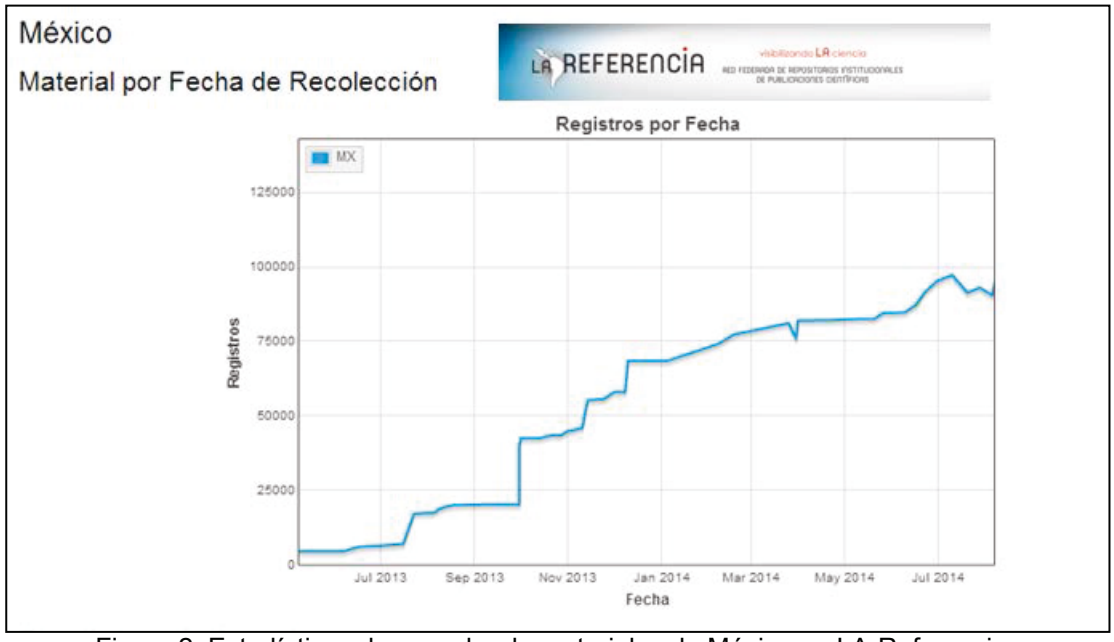

Figura 2. Estadísticas de cosecha de materiales de México en LA-Referencia disponible en: http://www.lareferencia.info/vufind/Laref/Cosecha?iso=MX

REMERI fue la primera red nacional en cumplir con lineamientos técnicos específicos de LA-Referencia, incluyendo los términos "instname" y "reponame" y requerimientos para DRIVER. El porcentaje de transformación (adecuación) de metadatos por parte de LA-Referencia es de 0.1 %, siendo el nodo nacional con el menor número de transformaciones y que tiene el mayor porcentaje de aceptación de registros, equivalente al 99.2 % en agosto 2014.

RESULTADOS Y CONCLUSIONES

Después de más de dos años de trabajo, REMERI cuenta en este momento (agosto del 2014) con la incorporación de 69 repositorios de 37 Instituciones Mexicanas para un total de 208,695 documentos incluyendo artículos (24.7%) , tesis de licenciatura (15.6%), tesis de maestría (21.9%), tesis de doctorado(6%) e imágenes (25%). En menor porcentaje se encuentran los videos, objetos de aprendizaje, trabajos recepcionales, capítulos de libros y libros.

Hasta el momento con INDIXE se han desarrollado 19 servidores de metadatos institucionales, incorporando 25 repositorios universitarios de REDALYC y mantenido el servidor de metadatos del proyecto interoperable y registrado en los directorios OpenDOAR y ROAR y con LA-Referencia. Además, se están desarrollado pruebas con INDIXE para gestionar redes de repositorios de producción cultural, revistas científicas y de tesis digitales.

Por otro lado, entre los principales retos a nivel nacional para el desarrollo y consolidación de repositorios digitales de acceso abierto, podemos mencionar los siguientes: La capacitación al personal responsable de la gestión de los Repositorios en tres competencias básicas: tecnológicas, informacionales y de comunicación; la estandarización y normalización de metadatos conforme a directrices internacionales de interoperabilidad; y, el establecimiento de políticas y mandatos de Acceso Abierto en las Instituciones Mexicanas de Educación Superior.

Finalmente, el proyecto REMERI que inicio con el desarrollo de un plan piloto, es ahora una Red de alcance nacional, que cumple con los estándares y directrices internacionales de interoperabilidad, siendo el primer proyecto mexicano en su tipo que además, está alineado a la estrategia nacional que está siendo coordinada por el CONACyT para el desarrollo del Repositorio Nacional de Producción Científica e Innovación, en el marco de la nueva Ley de Acceso Abierto de México.

REFERENCIAS BIBLIOGRÁFICAS

Abadal, E. (2012). *Acceso abierto a la ciencia*. Editorial UOC. Recuperado el 18 de febrero de 2014, desde: http://diposit.ub.edu/dspace/bitstream/2445/24542/1/262142.pdf

Alonso, J., Subirats I. y Martínez, M. L. (2008*).* Informe APEI sobre acceso abierto*.* Gijón: Asociación Profesional de Especialistas en Información. Recuperado el 9 de diciembre de 2013, desde: http://eprints.rclis.org/12507/

Bustos G. A., y Fernández P. A. (2008). *Directrices para la creación de repositorios institucionales en universidades y organizaciones de educación superior.* Recuperado el 13 de diciembre de 2013, desde: http://repository.urosario.edu.co/handle/10336/223

Budapest Open Access Initiative (2002). Recuperado 28 de enero de 2014, desde: http://bibliotecnica.upc.es/rebiun/BOAI.pdf

Crow, R. (2002). The Case for Institutional Repositories: A SPARC Position Paper. Recuperado el 3 de diciembre de 2013, desde: http://works.bepress.com/ir_research/7

Lynch, C. (2003). Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age ARL: A Bimonthly Report, no. 226 (February 2003). Recuperado el 30 de noviembre de 2013, desde: http://www.arl.org/storage/documents/publications/arl-br-226.pdf

Suber (2012). Open Access. Cambridge, Massachusetts: The MIT Press Essential Knowledge Series. Londres, Inglaterra.